Connecting the Dots: Integrating Point Location Data into Spatial Network Analyses

Shuto Araki[1] and Aaron Bramson[1,2,3(✉)]

[1] GA Technologies, Inc., Roppongi 3-2-1, Minato-ku, Tokyo 106-6290, Japan
[2] RIKEN Center for Biosystems Dynamics Research, Laboratory for Symbolic Cognitive Development, Minatojima-Minamimachi 6-7-3, Chuo-ku, Kobe 650-0047, Japan
[3] Ghent University, Department of General Economics, Tweekerkenstraat 2, 9000 Ghent, Belgium
a_bramson@ga-tech.co.jp

Abstract. Transportation networks allow us to model flows of people and resources across geographic space, but the people and resources we wish to model are often not natively tied to our networks. Instead, they can occur as point data (such as store, train station, and domicile locations) and/or grid data (such as socio-economic and aggregate area data). Here we present a set of methods to integrate point data into an augmented transportation network. This method facilitates analyses of temporo-spatial measures (such as accessibility scores) using only efficient breadth first search algorithms. We demonstrate the approach by calculating walkability scores for the train stations within the central Tokyo area.

Keywords: Spatial networks · Transportation networks · Data integration · Accessibility · Walkability

1 Introduction

Performing transportation network analyses typically requires data from multiple sources, and some of them may not fit into the network as attributes of existing nodes or edges. Both point and grid data are examples of data requiring additional integration steps. If one considers these points as additional nodes, one needs to consider how these point nodes should be related to/from other nodes in the network and how those configurations affect network features.

This paper presents a refined set of methods for integrating additional location point data into a transportation network in ways that improve accessibility scoring (among other things). Using the central Tokyo metropolitan area as an example, we demonstrate the effectiveness of our integration method by ranking railway stations by their 'walkability,' a measure reflecting the degree to which

R. M. Benito et al. (Eds.): COMPLEX NETWORKS 2020, SCI 944, pp. 193–205, 2021.
https://doi.org/10.1007/978-3-030-65351-4_16

surrounding amenities are reachable by foot. Our method produces a more realistic and personalizable metric of accessibility to surrounding stores than a simple count of nearby stores.

The integration methods described here are useful beyond the calculation of walkability scores of train stations. We chose train stations because their locations are publicly available and the Tokyo urban area is highly train-centric: half of all transportation uses the rail system and urban development is planned around stations [2,3]. Considering this, it is natural to focus on these stations to assess and compare accessibility.

2 Data Sets

Network Data. Our base network data is the road network for the Tokyo area from Open Street Map (OSM) [13]. The OSM road network includes nodes for all intersections as well as nodes to capture the curvature of the roads with straight edges. We isolate the largest connected component (98% of the nodes); this action removes remote islands, many pedestrian walkways, some access roads (e.g., within amusement parks and gardens), etc. Ideally we would use the footpath network data that includes sidewalks, pedestrian bridges, multi-use paths, greenways, etc. to more accurately determine accessibility via walking [4], however such data is not reliably available for Tokyo at present. Using the coordinates of the nodes, we generate the length of each edge using Haversine distance.

Point Data. Our store data comes from NTT Townpage [12], a private data service that provides lists of stores and other entities by category based on phone numbers. For the current demonstration, we limit our analysis to within Toyko's 23 Wards (central Tokyo) and to establishments within the following categories: variety store, hobby, travel, restaurant, cafe, bar, supermarket, convenience store, hospital, drugstore, laundry, public bath, spa, hotel, sport shop, sporting, cram school, nursery school, religion, and areas of concern (such as gambling establishments). For simplicity we refer to all these establishments as 'stores.' Note that because store *locations* are considered as nodes, there can be multiple stores in one store node.

In order to assign walkability scores to train and subway stations in Sect. 4, we also need to incorporate the locations of those railway stations. We use data from [11] for the names, longitude, and latitude of the stations, and augment this with the locations of subway entrances and station turnstiles from OSM [13].

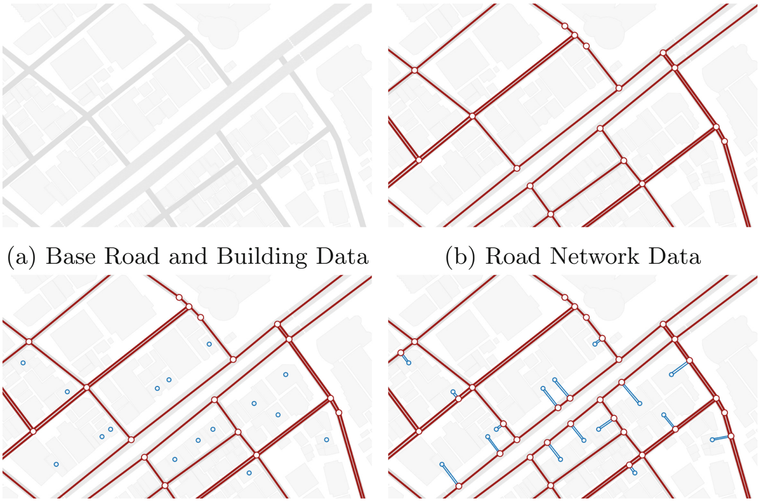
3 Integrated Analysis

An analysis using the integrated datasets requires two components: network construction and network traversal. First we describe the procedure to augment the road network with additional point data and edges, then we describe changes to network traversals incorporating mobility considerations.

3.1 Connecting the Road Network to Points

Both store and station/exit nodes require integration with the road network, and our method differs between them.

Integrating the Stores to the Road Network. The first step of the multimodal network construction is to integrate the stores to the road network. The general process is shown in Fig. 1; each store's building location is connected to the road network at its closest point. Specifically, for each store location, there needs to be an edge between the store and a newly created node at the closest point along its closest road edge. Because the distance is minimized, the road edge and store access edge are perpendicular. When the closest element of the road network is a node rather than an edge, we connect the store directly to that road network node. Although this method emulates how one enters the stores, the actual store entrance may be different from the closest road network edge/node.



(a) Base Road and Building Data

(b) Road Network Data

(c) Store Location Point Data

(d) Road Network Extended to Stores

Fig. 1. Steps to connect the store point data to the road network data to create the store-augmented multimodal transportation network.

In order to minimize the computational time to find the nearest edge, the Sort-Tile-Recursive (STR) tree data structure [9] is used to identify the closest edge from each store location. First, we construct an STR tree by treating the network edges as line geometry objects. Second, we convert each store node into a circle geometry object by adding a 125 m radius buffer. In this way, we can query the STR tree for all edges that overlap or intersect with the store's circle. From this set of closest edge candidates, we use a binary heap data structure (min-heap) to track the actual closest edge. Because the distance calculation is computationally expensive compared to the STR tree query, this two-step method saves considerable time compared to checking the distance between each store to every edge.

Setting the Coordinate Reference System. One caveat to this method is that the distance between a store and its nearby edges can be calculated using standard Euclidean distance only if the Coordinate Reference System (CRS) approximates a Euclidean space around the area of interest. Since the earth is nearly spherical, analyzing geospatial data on a flat plane requires a projection, and any projection of a 3D surface onto a 2D surface comes with some amount of distortion. The Mercator projection, for example, is often used for mapping because it preserves shapes and angles, but it heavily distorts areas and distances as you get further away from the reference point. Figure 2 shows the difference between the Web Mercator projection (epsg:3857 in red) and a distance-preserving CRS (`+proj=eqc +lat_0=35.6825 +lon_0=139.7671 +units=m` in blue). Since the visualization engine (Kepler.gl [7]) uses the Web Mercator projection, it visualizes edges that are connected to minimize distance using distance-preserving CRS as crooked (i.e., they do not appear perpendicular). If we calculate minimum distance using the Web Mercator projection, the edges appear perpendicular and correct to our eyes; however, the distances have been distorted. As seen in Fig. 2, this small discrepancy in CRS settings can cause some store nodes to be connected to a different edge. Misidentification of the nearest edge from a certain store could therefore influence the accuracy in assessing reachability of the store, and we utilize the distance-preserving CRS throughout.

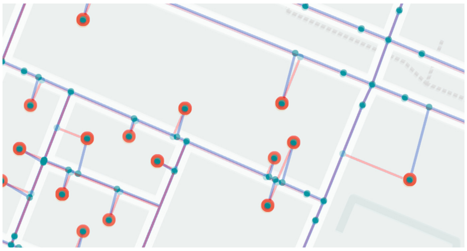


Fig. 2. The choice of coordinate reference system (CRS) impacts the calculation of distances and visualization of edges extended to stores. Red edges look perpendicular when calculated using the same Web Mercator projection used for visualization, but the blue edges are the shortest when calculating with a distance-preserving CRS.

Integrating the Station Exits to the Road Network. Our approach to integrating rail station data differs from the method used for stores. While stores are single point data representing the center of the appropriate building, stations are often large and multiply connected structures. There are three kinds of rail systems: subway, surface rail, and trams (streetcars). Ideally we would connect the entrances/exits of the stations to the closest point on the road network, and then connect the entrances to a station's main location point. In the OSM data, nearly all subway stations have accurate exit locations, but most surface and tram stations have only one exit point per station (and it's not an actual exit point). This limitation in our data requires us to approximate how the stations are connected to the road network.

Rather than handle stations on a case-by-case basis, we decided to create an adaptive rule that can be parsimoniously applied in order to maintain generality and hence applicability beyond Tokyo. First, each exit node is connected to the station node to which it is closest. In some case this differs from the station for which that exit officially corresponds, but this is a reasonable approximation considering the high level of interconnectedness within stations and the practical implications for access.

Station exit nodes are connected to all road nodes within a certain, but adaptive, distance around them. The radius of connectivity is determined in the following manner: Starting with $r = 10$, if there is at least one road node within a circle with radius of r meters, connect to all road nodes within a larger circle of radius $r + 10$ m. Else, increment r by 10 and repeat the process. This results in a station-augmented network shown in Fig. 3 (note that we use the road network after it has been augmented by the store access nodes).

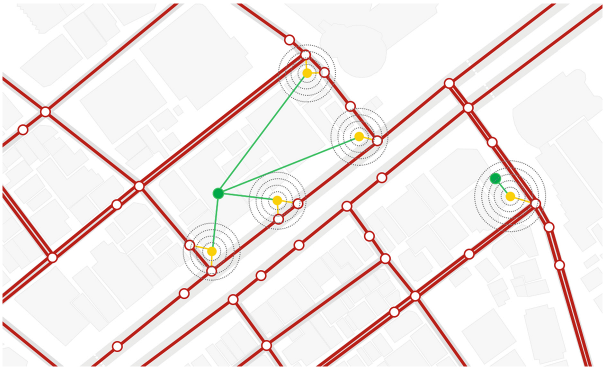


Fig. 3. Stations (green) are connected to exits (yellow) which are in turn connected to their surrounding road nodes (red). Using increasingly larger circles ensures at least one connection, but may also include additional road nodes within a similar distance (Color figure online).

In practice this method overconnects the exit nodes to the dense road network, resulting in redundant edges as can be seen in Fig. 4 (yellow lines). However, a stricter rule suffers from underconnection; specifically, stations are often connected only on one side and this underestimates their accessibility. Although extra links worsen the performance of network traversals, they have a negligible effect on accessibility measurements, which we consider to be of greater importance. In the future we will explore ways to reduce the edge redundancy in a parsimonious and generalizable manner.

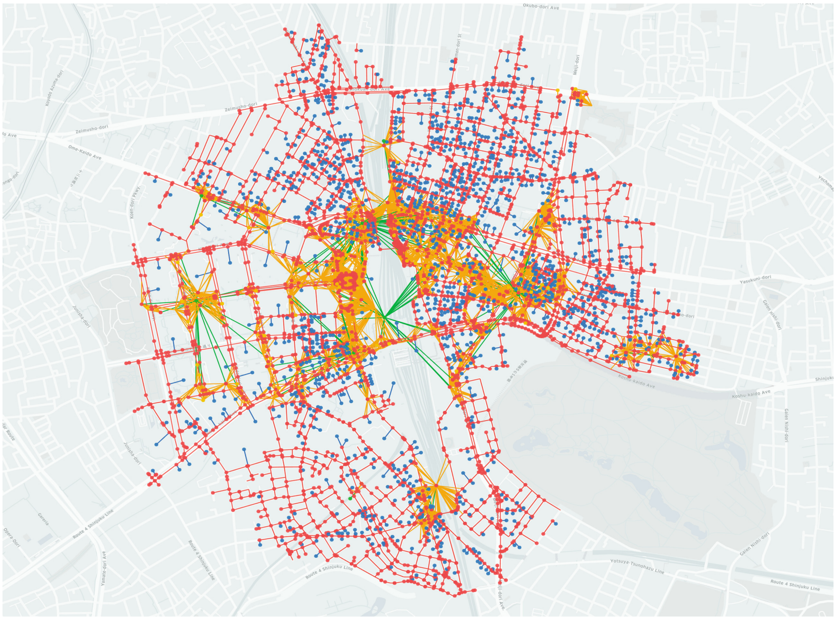


Fig. 4. Geospatial network diagram showing the subgraph induced by traversing 15 min along the integrated network from the node for Shinjuku station. Green station nodes are linked to yellow/red exit nodes via green links; yellow exit nodes are linked to red road nodes via yellow links; red road nodes are linked to each other via red links and to blue store nodes via blue links. Note that the Kepler.gl visualization engine introduces artifacts (missing nodes, wrong colors, etc.) not present in the data (Color figure online).

3.2 Integrated Network Traversals

The simplest measure of accessibility is the number of stores within a radius of focal point. A simplistic network-based approach uses the number of stores within a distance to a reachable road segment. However, our fully integrated network approach allows us to precisely measure the time required to traverse any origin-destination path using standard breadth first search algorithms.

In this paper, the accessibility score of a station is the time-weighted total number of stores reachable from that station. Each store node j has $m_j \geq 1$

stores located there. The contribution ω_j of store node j to the accessibility of station node i is time-discounted using a cosine-based function that reaches zero at T (shown in Eq. 1). We chose this functional form because it allows us to control the willingness to walk via the λ parameter to emulate objectively-measured moderate and vigorous physical activity (MVPA) data by age cohort [5,17]. Lower λ values correspond to those who prefer shorter distances while larger λ values delay the reduction in score contribution. Obviously, the cosine function rebounds after T, so we prune $t > T$. The weight value by traversal time for three values of λ are shown in Fig. 5. Further adjustments to T and λ, or alternative functional forms, can capture other means of transportation.

$$\omega_j = \frac{m_j}{2}\left(1 + \cos\frac{\pi t_{ij}^{\lambda}}{T^{\lambda}}\right) \tag{1}$$

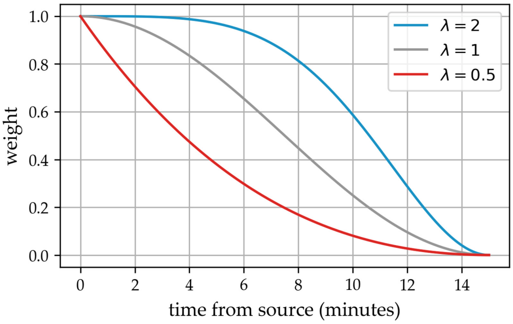


Fig. 5. A plot of the function for discounting the number of reachable establishments by the time needed to reach them using $T = 15$ (15 min by foot; 1250 m at 5 kph). Higher λ values delay the score reduction.

4 Demonstration via Walkability Scores

The focus of this work is presenting the refined physical network integration methods, which are generally applicable across mobility and accessibility research. By integrating a fine-grained road or footpath network augmented with access nodes and links to points of interest one can 1) determine best paths based on multiple criteria, 2) score and classify regions based on network features, 3) evaluate the impact of construction plans, and 4) assess various other social and transportation issues. As already noted, it is especially useful when scoring places based on their accessibility on the network. We demonstrate its effectiveness using a simple walkabilty scoring application.

4.1 Accessibility Scoring

We use the term ‘accessibility’ as an umbrella concept that includes any assessment of the ability to reach/use surrounding resources, broadly construed. Various measures of accessibility have been developed over the years, and applications have ranged from access to job, access to other people, access to food shopping, etc. [1,10,14]. Early development of accessibility research focused on efficiency and energy consumption [5] while recent ones have focused more on personalized metrics. For example, Quercia et al. [15] identifies the paths between points in London that are more beautiful, quiet, and happy. Using data about slopes, steps, ramps, elevators, etc. one can determine how well people with specific disabilities can access a location as well. All of these count as accessibility scoring, and in this paper, we refer to accessibility via walking as ‘walkability’.

4.2 Walkability Scores

Driven by a desire to promote exercise and reduce carbon emissions from vehicles, there has been a recent boom in research on walkability. One often used measure of walkability in particular is *Walk Score*® [18], which is focused on North America but partially validated for Japan by [8]. That measure’s details are not public, so we can’t reproduce them for comparison; however, it seems to simply count the number of establishments reachable from nodes of the large-scale road network using decreasingly sized circular buffers based on the time to that node. The *Walkability Index* of [16] uses a uniform buffer on the fine-grained network, but includes other considerations such as diversity of establishments. No available method leverages an integrated network to discount the contribution of further establishments.

We compare five different walkability measures. The first one is a baseline that counts of stores within 1250 m (the distance an average person can walk in 15 min) from each station. The second one uses Dijkstra’s algorithm on the integrated network to determine the number of accessible stores, but no time-weighting is applied. The third, fourth, and fifth measures take the same results from Dijkstra’s algorithm, but use the discount function shown in Eq. 1 with $T = 15$ and three different λ values (2.0, 1.0, and 0.5) to weight the results.

As explained in Sect. 3.2, these different λ values are designed to model one’s willingness to walk. For example, with a walkability score of $\lambda = 0.5$, a station with stores very close to the station could get a higher score than another station with twice as many stores within the 1250 m circle, but all more than 5 min away because at $t = 5$ the station counts are already discounted to around 40% as shown in Fig. 5.

Case Analysis. Although there is no “true” walkability level against which to measure accuracy, we do find that these different types of walkability scores reveal interesting differences in what they are measuring. Table 1 shows the top 15 stations ranked by the five measures. We recognize that many readers are not familiar with the areas of Tokyo, so we will explain the kinds of insights our method reveals using a few example stations.

Table 1. Comparison of the highest walkability stations by score.

rank	in 1250m radius	unweighted	$\lambda = 2$	$\lambda = 1$	$\lambda = 0.5$
1	Yurakucho (6064)	Ginza (4749)	Ginza (3066)	Ginza (2023)	Ginza (980)
2	Ginza (6023)	Ginza Icchome (4327)	Shinbashi (2833)	Shinbashi (1880)	Shinbashi (934)
3	Ginza Icchome (6015)	Hibiya (4201)	Shinjuku Nishiguchi (2525)	Shinjuku Nishiguchi (1669)	Shinjuku Nishiguchi (838)
4	Kyobashi (5707)	Shinbashi (4180)	Shinjuku (2478)	Shinjuku Sanchome (1602)	Shinjuku Sanchome (821)
5	Hibiya (5548)	Yurakucho (4115)	Uchisaiwai (2427)	Shinjuku (1560)	Shinjuku (718)
6	Higashi Ginza (5389)	Higashi Ginza (3987)	Shinjuku Sanchome (2409)	Uchisaiwai (1450)	Ikebukuro (651)
7	Shinbashi (5320)	Uchisaiwai (3891)	Ginza Icchome (2261)	Seibu Shinjuku (1367)	Seibu Shinjuku (650)
8	Takaracho (5309)	Shinjuku Nishiguchi (3667)	Yurakucho (2194)	Ginza Icchome (1324)	Uchisaiwai (646)
9	Shiodome (5210)	Shinjuku (3658)	Seibu Shinjuku (2172)	Yurakucho (1269)	Ginza Icchome (621)
10	Tsukiji Market (4977)	Shiodome (3603)	Higashi Ginza (2167)	Higashi Ginza (1238)	Shinsen Shinjuku (583)
11	Uchisaiwai (4669)	Shinjuku Sanchome (3579)	Shinsen Shinjuku (2036)	Ikebukuro (1232)	Yurakucho (568)
12	Nihonbashi (4369)	Seibu Shinjuku (3501)	Hibiya (1797)	Shinsen Shinjuku (1218)	Ueno Hirokoji (558)
13	Onarimon (4232)	Shinsen Shinjuku (3464)	Shiodome (1723)	Shibuya (1074)	Higashi Ginza (556)
14	Tokyo (4207)	Kyobashi (3190)	Ikebukuro (1698)	Ueno Okachimachi (1017)	Ueno Okachimachi (547)
15	Shinjuku Nishiguchi (4179)	Takaracho (2892)	Kyobashi (1562)	Ueno Hirokoji (998)	Shibuya (547)

The most obvious pattern is that the Ginza metro station dominates this ranking. Ginza is known for its massive shopping streets, eateries, and entertainment venues; thus it is not surprising to see that Ginza and its nearby stations (Shinbashi and Yurakucho) are consistently ranked near the top of the lists. Perhaps more interesting is that many stations surrounding Ginza (Ginza Icchome, Kyobashi, Hibiya, Higashi Ginza) fill the top spots of the in-radius ranks, but are pushed down further and further as we move to unweighted, and increasingly strict discounting.

We now take a closer look at one of those surrounding stations: Hibiya metro station. It is ranked 5th and 3rd in the circle baseline and unweighted approach, respectively; however, its rank drops significantly as the discounting is applied (12 → 16 → 20). The reason for this is clear when looking at a map: one can reach Yurakucho within 10 min and almost to Ginza within 15 so its reach includes many of the surrounding larger shopping streets. But there are not many stores around Hibiya station itself: one corner has the sprawling Imperial Palace and another the famous Hibiya Park. This result demonstrates the need for a weighted walkability score; failing to discount the contribution of further stores falsely promotes locations on the fringe of major shopping districts while downplaying the convenience of locations in the middle of smaller shopping areas.

Surprisingly, Shinjuku station, the busiest station in the world [6], also famous for its huge shopping and entertainment areas, only appears at rank 15 in the circle baseline approach, and even then it is a satellite station rather than the main one. The reason is Shinjuku station's immense size: the station itself is hundreds of meters long and wide, so other stations (especially subway stations that have practically zero footprint) have more stores that are physically close. Those stations benefit from the distance one can travel in 15 min and the proportion of the area that supports having stores. One can see Shinjuku (as well as Shinjuku Nishiguchi, Shinjuku Sanchome, Seibu Shinujuku, and Shinsen Shinjuku) rising up the ranking as λ gets smaller. Those 'walkable' stations that were ranked high in the circle baseline approach rapidly fell from the ranking because their scores got significantly discounted.

Ikebukro station exhibits a similar trend, but more drastic. Ikebukro is a secondary city center with many stores (though not as many as Ginza or Shinjuku) nearby the station, but as a more recent development they do not sprawl out into surrounding territory. It did not rank high in either baseline (67th and 34th), but because it has a somewhat large number of stores focused around the station it reaches ranks $14 \rightarrow 11 \rightarrow 6$ as λ decreases. This is an important case because anybody familiar with Tokyo would agree that Ikebukuro is a major and convenient shopping and entertainment hub, but the unweighted measures could not reveal this characteristic.

Similarity Analysis. Although an analysis of specific stations allows us to compare the resulting walkability scores with our intuitions, Fig. 6 shows the similarities between each pair of measures using the Kendall rank correlations. This statistic takes two ordered lists and computes the number of pairs in the same order, minus the number of pairs in a different order, and divides by the number of possible pairs. It informs us how similarly two lists of the same items are ranked.

Note that the circle baseline approach is most similar to the unweighted approach and becomes less similar as λ decreases. This result is not surprising considering how the discount function heavily penalizes stores further away. One can also observe the high similarities among the four network-based measures. The fact that the unweighted approach is more similar to the case with $\lambda = 2$ than it is to the circle baseline approach suggests that there exist some distinct features that the network traversal was able to extract (i.e., stores that are within 1250 m but not actually reachable within 15 min due to circuity of the network and barriers such as rivers, railways, and highways) that are more important than the weighting. Although these similarity results are unsurprising, it is reassuring to get a confirmation of the intuitive relationships among these measures.

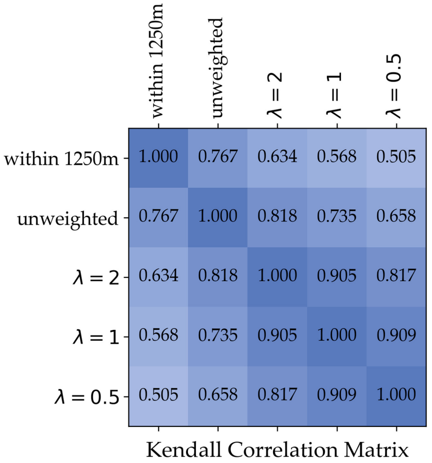


Fig. 6. Pairwise comparisons of the Kendall τ coefficients.

5 Conclusions and Future Work

Using the fine-grained road network data facilitates the discovery of accurate paths and therefore accurate traversal times. Augmenting this network to parsimoniously integrate access edges to points of interest (such as train stations and stores) allows us to calculate times from an origin to each potential destination using efficient network search algorithms. After describing our novel methods for capturing this physical network system, we presented a comparison of walkability scores showing the importance of network-based assessments and discounting establishments that are further away. We also demonstrated how varying the time-weighting parameter can capture differences in accessibility for different populations, such as the elderly or disabled.

Based on this preliminary analysis, the integrated network achieves more believable scores compared to the circle baseline approach because walking paths in Tokyo are often meandering and complicated. To get from point A to point B, there rarely exits a straight path and therefore the circle baseline approach overestimates the number of realistically reachable stores. However, applying a discount function to the circle approach might be a good approximation of the integrated analysis because the further stores would get heavily discounted scores. We are currently investigating this approach for basic scoring.

Although our network augmenting methodology produces more accurate paths, traversal times, and walkability scores, we recognize that accessibility measures that only include the time to places of business offer a narrow view of walkability. Rather than just focus on the degree to which people can get their shopping done on foot, one might also consider how pleasant an area is to walk through [15]. Including locations such as parks, gardens, riverside paths, scenic views, etc. offers a score of walk-worthiness. We could produce different measures for the various populations, interests, and purposes, and then generate a walkability score that combines these measures.

For all these purposes and interests, the paths must be further analyzed beyond just traversal times. By incorporating building heights and footprints we can characterize neighborhoods by their openness. Data on green areas such as road-side trees and grassy medians is also clearly relevant. Typical noise and traffic levels can also be used to improve our assessment of walkability. Perhaps the most important factor needing inclusion is the slopes of road segments and a measure of the traversal effort. All these, along with parameterizations for bicycles, wheelchairs, and other mobility factors, are included in the walkability index we are developing based on the network methodology presented in this paper.

References

1. Biazzo, I., Monechi, B., Loreto, V.: Universal scores for accessibility and inequalities in urban areas. arXiv preprint arXiv:1810.03017 (2018)
2. Calimente, J.: Rail integrated communities in Tokyo. J. Trans. Land Use **5**(1), 19–32 (2012)
3. Chorus, P., Bertolini, L.: An application of the node-place model to explore the spatial development dynamics of station areas in Tokyo. J. Transp. Land Use **4**(1), 45–58 (2011)
4. Ellis, G., Hunter, R., Tully, M.A., Donnelly, M., Kelleher, L., Kee, F.: Connectivity and physical activity: using footpath networks to measure the walkability of built environments. Environ. Plan. B: Plan. Des. **43**(1), 130–151 (2016)
5. Frank, L., Ulmer, J., Lerner, M.: Enhancing walk score's ability to predict physical activity and active transportation. In: Active Living Research Annual Conference, San Diego, CA Retrieved from (2013). http://activelivingresearch.org/sites/default/files/2013_Bike-WalkScore_Frank. pdf
6. Guinness World Records: Busiest station (2018). https://www.guinnessworldrecords.com/world-records/busiest-station
7. keplergl: kepler.gl, August 2020. https://github.com/keplergl/kepler.gl
8. Koohsari, M.J., Sugiyama, T., Hanibuchi, T., Shibata, A., Ishii, K., Liao, Y., Oka, K.: Validity of walk score$^{®}$ as a measure of neighborhood walkability in japan. Prevent. Med. Rep. **9**, 114–117 (2018)
9. Leutenegger, S.T., Lopez, M.A., Edgington, J.: STR: A simple and efficient algorithm for r-tree packing. In: Proceedings 13th International Conference on Data Engineering, pp. 497–506. IEEE (1997)
10. Levinson, D.: Network structure and city size. PloS one **7**(1), e29721 (2012)
11. Lüthy, M.: japan-train-data, May 2017. https://github.com/adieuadieu/japan-train-data
12. NTT Townpage Inc.: Townpage Database. Proprietary Dataset, July 2019
13. OpenStreetMap Contributors: Planet dump retrieved from planet.osm.org (2019). www.openstreetmap.org
14. Páez, A., Scott, D.M., Morency, C.: Measuring accessibility: positive and normative implementations of various accessibility indicators. J. Transp. Geo. **25**, 141–153 (2012)

15. Quercia, D., Schifanella, R., Aiello, L.M.: The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In: Proceedings of the 25th ACM conference on Hypertext and social media, pp. 116–125 (2014)
16. Shimizu, C., Baba, H., Kawase, T., Matsunawa, N.: Walkability and real estate value: Development of walkability index. Online, June 2020. http://www.csis.u-tokyo.ac.jp/wp-content/uploads/2020/06/163.pdf
17. Trost, S.G., Pate, R.R., Sallis, J.F., Freedson, P.S., Taylor, W.C., Dowda, M., Sirard, J.: Age and gender differences in objectively measured physical activity in youth. Med. Sci. Sports Exerc. **34**(2), 350–355 (2002)
18. Walk Score: Walk Score® (2020). https://www.walkscore.com/